Iterative Fragmentation Improves the Detection of ChIP-seq Peaks for Inactive Histone Marks

Miklós Laczik[1,2], Jan Hendrickx[2], Anne-Clémence Veillard[2], Mustafa Tammoh[2], Sarah Marzi[3] and Dominique Poncelet[2]

[1]Doctorate Student, Doctoral College of Agronomy and Bioengineering, Gembloux Agro-Biotech, University of Liège, Liège, Belgium. [2]Researcher, R&D Epigenetics Department of Diagenode SA, Liège, Belgium. [3]Doctorate Student, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.

ABSTRACT: As chromatin immunoprecipitation (ChIP) sequencing is becoming the dominant technique for studying chromatin modifications, new protocols surface to improve the method. Bioinformatics is also essential to analyze and understand the results, and precise analysis helps us to identify the effects of protocol optimizations. We applied iterative sonication – sending the fragmented DNA after ChIP through additional round(s) of shearing – to a number of samples, testing the effects on different histone marks, aiming to uncover potential benefits of inactive histone marks specifically. We developed an analysis pipeline that utilizes our unique, enrichment-type specific approach to peak calling. With the help of this pipeline, we managed to accurately describe the advantages and disadvantages of the iterative refragmentation technique, and we successfully identified possible fields for its applications, where it enhances the results greatly. In addition to the resonication protocol description, we provide guidelines for peak calling optimization and a freely implementable pipeline for data analysis.

KEYWORDS: chromatin, heterochromatin, histone marks, ChIP, ChIP-seq, sonication, bioinformatics, peak calling

CITATION: Laczik et al. Iterative Fragmentation Improves the Detection of ChIP-seq Peaks for Inactive Histone Marks. *Bioinformatics and Biology Insights* 2016:10 209–224 doi: 10.4137/BBI.S40628.

TYPE: Perspective

RECEIVED: July 25, 2016. **RESUBMITTED:** August 23, 2016. **ACCEPTED FOR PUBLICATION:** August 28, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 1658 words, excluding any confidential comments to the academic editor.

FUNDING: This study was financed by Diagenode SA and the DisChrom (Chromatin diseases: from basic mechanisms to therapy) project under the grant agreement PTN-GA-2009-238242. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: smice@email.com

Introduction

One of the focal points in epigenetics research is the study of chromatin, the complex of deoxyribonucleic acid (DNA) and bound proteins. Chromatin is not homogenous, it is dynamically changing, and its structure is heavily influenced by the modifications of the histone proteins that the DNA is bound to. In turn, the chromatin structure has substantial effects on the DNA, making some genes more accessible for transcription while effectively silencing others.[1] Errors in this delicately balanced system lead to severe diseases, from various types of cancer[2–4] to specific chromatin diseases such as the α-thalassemia/mental retardation syndrome, X-linked (ATR-X syndrome), the immunodeficiency, centromere instability and facial anomalies syndrome (ICF syndrome) and the Rett-syndrome.[5] Therefore, it is important to study and understand the chromatin, its dynamics, and the role of the underlying histone modifications.

Chromatin immunoprecipitation (ChIP) is a widespread, established method for researching histone modifications. The DNA and the bound proteins are fixated with a crosslinking agent (eg, formaldehyde), then fragmented (with ultrasonication, enzyme digestion, hydroshearing, or other methods), and the random fragments carrying a certain histone mark are captured by antibodies specific to that histone modification. After decrosslinking, the positions of the histone modification of interest across the genome can be determined by analyzing the captured DNA. For the latter step, several detection methods exist, such as polymerase chain reaction (PCR), microarray hybridization, and massively parallel sequencing. While all methods have their unique advantages and disadvantages, the sequencing-based method (ChIP-seq) has become the prevailing one because it gives high-resolution quantitative genome-wide data without the drawbacks of the microarray technology, such as the lower genome coverage, lower resolution, limited availability of platforms for different species, and the requirement of a larger quantity of sample material. At the same time, the sequencing costs are decreasing steadily, while throughput is increasing, thanks to novel sequencing machines, reagents, and protocols, making ChIP-seq an ever more cost-effective choice.[6]

Several protocols exist for ChIP-seq, focusing on optimization for different sample types, sample amounts, or specific library preparation methods.[7–9] Many of these protocols contain sonication in a high-intensity bath sonicator for

the fragmentation of the chromatin, as it produces random fragments reliably with low bias and low chances of contamination and sample degradation, as opposed to other methods, such as enzymatic digestion, nebulization, hydrodynamic shearing, and tip sonication.[10] In addition, bath sonicators can usually process several samples at the same time, which greatly reduces the sample preparation time. We intended to explore the benefits of modifying the standard protocol by additional rounds of sonication, or to be more precise, reshearing the already sonicated and immunoprecipitated DNA after decrosslinking. A similar double fragmentation method has been described previously,[11] where the authors described how reshearing can be used to increase the chromatin yield and conduct successful ChIP-seq experiments from a low amount of sample, even from sub-nanogram amounts of immunoprecipitated material. They focused on transcription factors and an active histone mark (H3K4me3) that generates peaks exhibiting similar characteristics to those of transcription factors. In such cases, the collection of an adequate amount of immunoprecipitated DNA can be problematic, as transcription factors occupy only a diminutive fraction of the genome; thus, it can be pivotal to include as much of the captured fragments in the library generation as possible. However, conventional ChIP-seq protocols with size selection severely limit the availability of fragments for sequencing.

While we can confirm their findings (we routinely experienced a 5×–10× increase in DNA yield on average when we applied the reshearing technique in similar experiments, ie, ChIP-seq with active histone marks that generate narrow peaks), more importantly, we identified another field of application where this technique could prove essential: ChIP-seq experiments with inactive histone marks. In the above-mentioned study, they did not test the method on inactive histone marks, nor histone marks that yield broad regions of enrichment. Our hypothesis was that for inactive histone marks, the single shearing and size selection causes the loss of important material, because these marks tend to generate larger fragments, which are outside of the ideal fragment size range for ChIP-seq. Inactive marks, such as H3K27me3, typically cover long, continuous regions and form heavily condensed (facultative) heterochromatin,[12,13] which in turn is more resistant to breaking.[10,14–16] As a consequence, these regions have a higher presence among the longer fragments after the initial shearing step. Compared to the method described by Mokry et al.[11], we also propose improvements, such as an iterative approach to the additional shearing rounds, and minor technical variations in the sonication device, sonication settings, and sonication solution among others. We optimized our method specifically for inactive, broad histone marks.

Another important source of difference from the study of Mokry et al is the bioinformatics. Nowadays, the analysis of the massive amounts of – often genome-wide – data would be unimaginable without proper bioinformatics. We also believe that bioinformatics should be highly application specific and the analyzer should always strive after avoiding the introduction of bias and artifacts in the data. The analysis of ChIP-seq data involves the detection and characterization of peaks, which are actually enrichments of reads in the ChIP-seq profile, which represent the sequenced fragments that the antibodies capture. We developed a custom analysis pipeline specifically for analyzing ChIP-seq data of histone marks, yielding either broad or narrow peak profiles, aiming to precisely monitor how reshearing the already sonicated immunoprecipitated DNA affects the detectability and peak characteristics of the histone marks we studied. This ChIP-seq analysis pipeline not only allows us to control several important aspects of the peaks, such as size, significance, and overlap ratios, but also enables the use of unique peak detection settings custom tailored to each mark, as opposed to the study of Mokry et al, where peaks were called with uniform settings for all samples. We also present our analysis pipeline in this study.

Though some ChIP-seq protocols mention an additional round of sonication of the chromatin if the fragment size range is incorrect,[17,18] we would like to point out that it is a different approach than the reshearing method we applied, as the former is meant to shear some of the larger chromatin fragments (instead of the decrosslinked DNA) further, while the smaller fragments are still discarded during size selection. The refragmentation method we used does not involve size selection as the majority of the fragments fell in the appropriate size range, and thus nearly the whole sample is preserved, allowing most of the fragments to proceed to sequencing. Furthermore, it is not the chromatin that is resonicated but the already eluted, purified, and immunoprecipitated DNA, eliminating such obstacles as the higher physical resistance to shearing forces of the heterochromatin.

Materials and Methods

Cell culturing and ChIP. We chose the well-established, widely known HeLa-S3 cell line as our test subject. The cells were cultured in Dulbecco's Modified Eagle's Medium + 10% fetal bovine serum at 37 °C under 5% CO_2. We used the Diagenode iDeal ChIP-seq kit for histones (catalogue number: C01010051) for the ChIP with the Diagenode polyclonal antibodies for the following targets: H3K4me1 (catalogue number: C15410194), H3K4me3 (catalogue number: C15410003–50), and H3K27me3 (catalogue number: C15410069). We followed the manual of the ChIP-kit (https://www.diagenode.com/files/products/kits/iDeal-chipseq-histones-x24-x100-manual.pdf) for cell collection, crosslinking, cell lysis, chromatin extraction and shearing, antibody binding, elution, decrosslinking, and DNA isolation. For the DNA purification, we used the Qiagen QIAquick PCR Purification kit (catalogue number: 28106), and we followed the instructions in its manual (https://www.qiagen.com/us/resources/download.aspx?id=3987caa6-ef28-4abd-927e-d5759d986658&lang=en). With the control samples,

we proceeded straight to library preparation after purification. For the refragmentation experiment, we added the following procedure before library preparation (following Step 4 in the iDeal ChIP-seq kit manual): we put 20 µL of immunoprecipitated DNA in elution buffer (buffer EB in the purification kit) into 100 µL capped tubes for additional rounds of shearing. We used these 100 µL tubes because we had found that the choice of the correct tube is crucial: large capacity, 15 mL or 1.5 mL tubes that are usually recommended for chromatin shearing compromised the sonication efficiency during the additional fragmentation of the DNA. We performed consecutive rounds of shearing; each round consisted of five cycles of 30 seconds ON and 30 seconds OFF, making a total of five minutes for each round. Between the reshearing rounds, a centrifuge was used to spin down the solutions. As different histone marks produce different characteristic DNA fragment length distributions, we ran a series of reshearing rounds and monitored the fragment size distribution after each round by gel electrophoresis on a 2100 Bioanalyzer (Agilent) to find the optimal amount of resonication needed for the sample in question. We found that two rounds of reshearing (2 × five cycles) were enough for H3K4me1 and H3K4me3, but for the longer fragments of H3K27me3, three rounds were needed to reach the optimal size range (refer to Fig. 1 for the shift in fragment size distribution as a result of the reshearing rounds), which is 200–800 bp based on Illumina recommendations (see the manual of the TruSeq kit below).

All the sonication steps in the protocol were performed on a Bioruptor Pico (Diagenode) sonication device. After the described reshearing step, we proceeded to library preparation with the resheared samples as well.

Library preparation and sequencing. For library preparation, we used the Illumina TruSeq ChIP Sample Prep Kit (catalogue number: IP-202-1012) and followed its manual

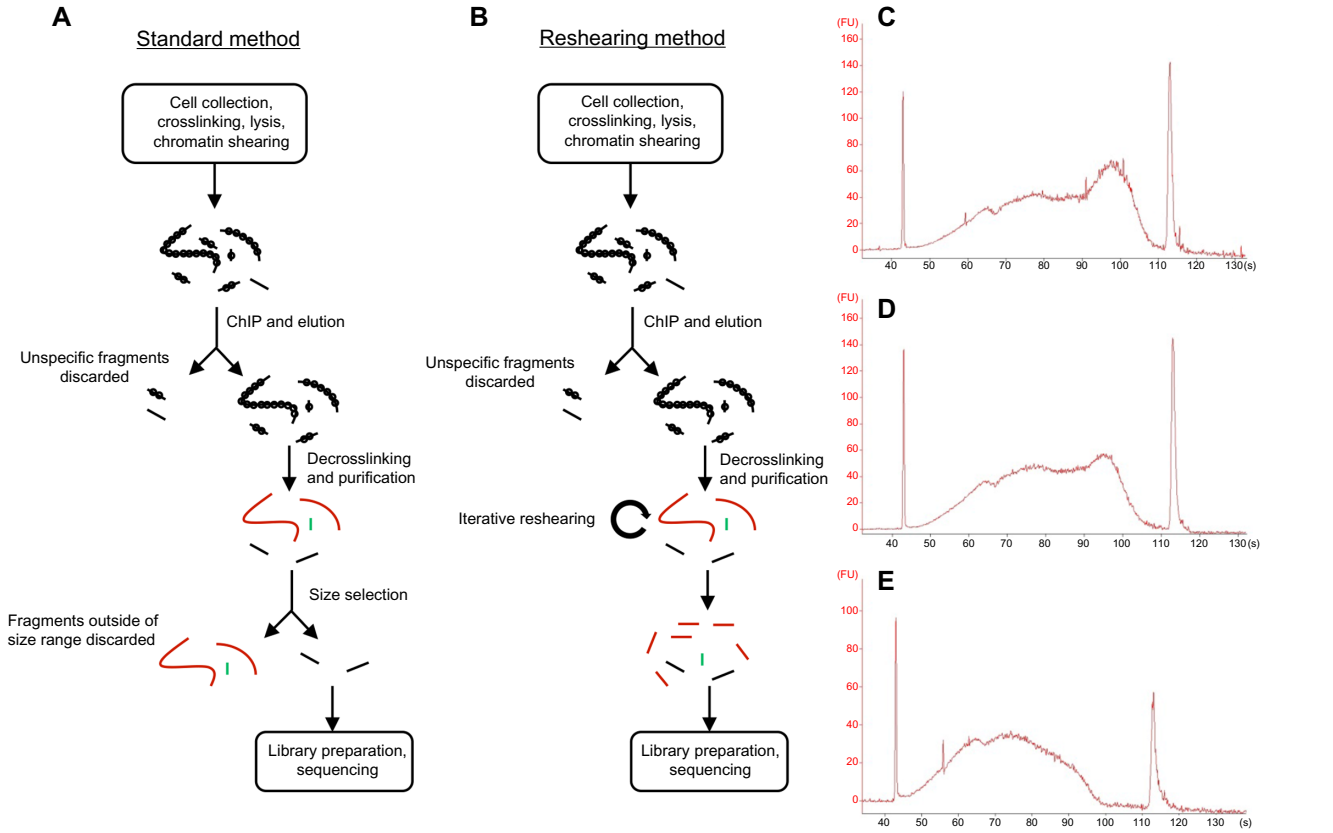

Figure 1. Overview and effect of the reshearing method. (**A**) The relevant steps of ChIP-seq sample preparation, using the traditional method. After the fragmentation of the protein, we have a mixture of fragments of different sizes. The ones that carry our protein of interest can be bound by immunoprecipitation. After the decrosslinking and purification step, we get the DNA fragments, where the ones over the optimal size range are shown in red, and the ones under the size range are in green for better visual interpretation. During the size selection before library preparation and sequencing, these fragments are discarded; thus, a significant amount of the sample is lost. (**B**) The reshearing method preserves the fragments that are out of the ideal size range. By doing additional rounds of sonication on the eluted DNA, the long fragments break up into shorter ones (see the fragments in red), which enables them to proceed to library preparation and sonication. Sample loss is reduced significantly. (**C**–**E**) Demonstrating the effect of the reshearing on the actual H3K27me3 sample. The fragment range distribution is measured by a 2100 Bioanalyzer, images were generated by its software provided by Agilent; the control marks are at 35 bp and 10380 bp. (**C**) The original fragment distribution, before the reshearing step. (**D**) The size distribution after two rounds of five cycles of reshearing. A reduction of the large peak in the large size range and a slight shift toward the smaller sizes is already visible. (**E**) The distribution after the third round of reshearing. Here the shift is already complete: the large fragments have disappeared and the middle short section of the size range is enlarged, showing that we have reached the desired size distribution.

(http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqchip/truseq-chip-sample-prep-guide-15023092-b.pdf) to prepare the libraries for Illumina sequencing for both the resheared and the control samples, except for the size selection step, which we skipped in the case of the resheared samples. Sequencing was performed by Fasteris SA on an Illumina HiSeq 2500 sequencer, with a setup to generate 50 bp long single reads.

Analysis pipeline. We developed a rigorous analysis pipeline aimed specifically at the analysis and quality control of ChIP-seq data of histone marks to fit our needs, as we have noticed a general shortage of analysis tools for peaks that are typical of histone marks. While transcription factors usually yield very narrow and high peaks, often narrower than 1000 bases, sometimes termed point-source peaks or punctuate marks, histone marks generate lower enrichments over a prolonged region, which could even span hundreds of kilobases – there is also a great variation in every dimension of these enrichments, dependent on the targeted histone mark. That is why ChIP-seq data of histone marks are more difficult to analyze, and there are less tools capable of it.

Our pipeline is a combination of acknowledged, peer-reviewed public software tools and our own algorithms developed in-house. For removing artifacts from the read set, such as adapter sequences and low-quality reads/bases, we apply TrimGalore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) (v0.4.0) with default settings, a wrapper tool for Cutadapt[19] (v1.8.1), followed by BWA[20] (v0.5.0) for aligning reads to the human genome (assembly version NCBI Build 36.1). We found the default settings of BWA to be adequate for our purposes, generating accurate alignments with a large portion of the reads (at least 80%), devoid of ambiguous and misalignments. The alignment and sequencing quality check was done by the FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (v0.11.2) software with default settings.

Before the construction of the analysis pipeline, we had tested a large number of peak calling tools to find the ones best suited for the wide range of enrichments found in a histone mark analysis. We identified them by searching the literature; we chose them based on their publications and also on the recommendations of relevant reviews and benchmarks.[21,22] We tested Sicer[23] (v1.1), MACS[24] (v1.4.2 and v2.0.9), Zinba[25] (v2.02.02), HiddenDomains[26] (v2.2), BroadPeak[27] (unversioned), and Rseg[28] (v0.4.4) on our data sets generated in-house (in the same way as we have described above the generation of control data sets, using the conventional ChIP-seq protocol, without refragmentation). We found that Sicer outperforms other peak callers in this regard. To determine the best settings for each histone mark, we conducted numerous experiments on data sets with the relevant histone marks. The peak calling principal of Sicer is windowing the genome, selecting candidate enrichment regions, merging the nearby ones, and filtering them by significance. The three key values among its options are thus the window size, gap size, and *E*-value. The essential settings for the histone marks included in this study are shown in Table 1.

Downstream analysis consisted of visual monitoring, generating various descriptive statistics, annotation, establishing peak profiles, and overlap analysis. For the visualization of various kinds of massively parallel sequencing data, we prefer the IGV browser from the Broad Institute[29] (v2.3.63). For the statistics, we used proprietary scripts written in R: our scripts calculate the peak numbers, the average of the peak widths, the average of the significance scores of the peaks, and the fraction of reads in peaks, as described by the ENCODE consortium.[21] The enrichments were annotated with the annotatePeaks.pl tool of the Homer software suit[30] (v4.7.2), using the –genomeOntology option and the appropriate annotation databases provided also by the Homer package, then from the annotation output, we use our scripts to generate ratios of peaks in three categories: promoters, coding regions, and gene-rich regions. The former two provide valuable insights for active marks, as they are often associated with genes and promoters, while the latter is important for all histone marks, generally they occur in gene-rich regions. The peak matching analysis (part of the overlap analysis) is done by the mergePeaks tool of the Homer software suit, set to report every overlap with the -d given option. Note that in the peak overlap analysis, we also examine the overlap of the top 40% of peaks, a method described by the ENCODE consortium.[21] Their criterion for two matching peak sets is that at least 80% of the top 40% of the peaks must overlap with the other data set. For this, our proprietary scripts generate the top 40% peak files by sorting the detected peaks based on their significance scores from most to least significant and writing out the top 40% of them into a new peak file. Our proprietary scripts also calculate the Pearson's coefficient of correlation based on the significance scores of the overlapping peaks. The other part of the overlap analysis, where we dissect the genome into windows and analyze the sample correlation window-by-window, as well as the average peak profile calculation, bias analysis, and visualization are performed entirely by proprietary scripts. Our scripts calculate the mean coverage in 100 bp windows and then based on them calculate the Pearson's coefficient of correlation across samples and produce scatterplots with the graphical tools provided by the base R package[31] (v3.0.1). Generating the average peak profiles is done by measuring the coverage in 100 bins

Table 1. Key values of Sicer we used for each peak calling. We determined these values empirically, using numerous test data sets with similar enrichment profiles.

HISTONE MARK	WINDOW SIZE	GAP SIZE	*E*-VALUE
H3K4me1	200	600	2000
H3K4me3	200	600	100
H3K27me3	5000	50000	5000

for each peak, then calculating the mean of each bin rank, and then displaying the data points per rank with the R plotting tools. Additional data processing steps were carried out with the help of SAMtools[32] (v0.1.19), BEDtools[33] (v2.22.1), and proprietary scripts: SAMtools was used to create BAM files from SAM files, sort them, and remove duplicates; BEDtools was used to create BED files from BAM files (needed by Sicer, it can only work with input alignments in BED format), and proprietary scripts were used to create a standard BED file as described on the UCSC website (http://genome.ucsc.edu/FAQ/FAQformat#format1) from the nonstandard peak files that Sicer outputs.

Note that although we implemented the custom part of our pipeline in R, this is not obligatory for it to work; it can be freely implemented on various other platforms, using other languages too.

Establishment of optimal parameters for peak calling. To find the optimal settings for peak calling, we used two approaches: we applied thorough visual inspection and we established a set of validated peaks whose detection we monitored.

For the first approach, we used the already mentioned IGV browser, where we could load both the reads (both as indexed alignments in BAM format or as coverage graph in bedGraph format, created with BEDtools from the alignment files) and the peaks (in BED format), then we were able to observe how well the detected peaks match the actual enrichments, whether the boundaries were recognized properly, or rather it displayed merging, partial detection, oversensitivity, or other peak calling problem. It was essential to go through several regions of the genome, also at several zoom levels to get a complete picture. Annotation proved to be helpful; for example, in the vicinity of domestic genes, we could always find prominent enrichment in case of active marks.

For the second approach, we used our in-house data sets generated with the highest standards and greatest care to achieve excellent quality data. For each of the three histone marks we studied in this study, we prepared five replicates with the respective antibodies using HeLa-S3 cells; we used the standard ChIP-seq protocol (without reshearing) as we described earlier for the generation of control data sets. For external validation, we have chosen the ENCODE data sets submitted by the Broad Institute, with the same cell type and histone marks (respective UCSC accession numbers of the H3K4me1, H3K4me3, and H3K27me3 data sets: wgEncodeEH001750, wgEncodeEH001017, and wgEncode EH001037). Sicer was run with the recommended settings in its documentation to identify enriched regions. (Note: for finding highly enriched sites, we do not need accurate peak calling, we only need to identify some genome windows with high coverage, so for this preliminary scan the default settings are more than adequate.) Our goal was to select 100 high fidelity enrichments, which are very consistent and well characterizable. Therefore, we identified the highest enrichment across all the data sets based on the preliminary scan and we checked if it is occurring in all the data sets, internal and external alike, examined it in the browser if it indeed looks like a good candidate, that is, it is not falling in dubiously mappable region (like the repeats near the centromere), and if it can be clearly defined. Annotation data also helped here, as we preferentially selected enrichments near enhancers and promoters in case of H3K4me1, promoters of active genes in case of H3K4me3, and known inactive regions and suppressed genes in case of H3K27me3. If the enrichment passed all criteria, then we included it in our data set, if not, then we moved on to the next best enrichment (unless it was already included), and we continued to do this iteratively until we collected 100 peaks, which had an extremely high chance of being true positives. Then we described these peaks in terms of their dimensions, and also we overviewed the whole enrichment profile along the genome to be assured that the selected peaks are not unique, they represent the general type of enrichments. Once we had this highly reliable peak set, we could run Sicer with a range of settings and monitor the results; we used BEDtools to compare the detected peaks and the actual enrichments as it can generate a number of useful statistics, that is, depth and breadth of coverage, histogram data for the coverage, and number of bases covered as well as the covered ratio of the total length for each peak.

To find the best settings, we first applied different settings with great increments, then we fine-tuned the settings with smaller increments, for example, for window size, we first used 200, 1000, and 5000, then if the shorter size seemed to be the superior, we went through window size settings 200, 400, 600, 800, and 1000.

Results

Experimental setup and analysis pipeline. Three different histone marks were tested with the iterative sonication method. Our primary goal was to monitor the effects of reshearing on the inactive mark H3K27me3, but we also tried out refragmentation on two active marks that produce narrow peaks: H3K4me1 and H3K4me3. We used the same cell type and same conditions (except for the different antibodies for the different histone marks for immunoprecipitation, and the rounds of reshearing, which were also dependent on the studied histone mark and the state of the captured chromatin, as we described earlier), and for each histone mark, we prepared a resheared sample and a control sample with the standard preparation method written in the manual of the ChIP-seq kit. For the sequencing, we found 50 bp single-end sequencing on an Illumina HiSeq machine to be ideal: this setup is adequate for obtaining high-quality data with well identifiable enrichments, while it avoids generating junk reads. The latter is important because with the additional shearing beside the optimal fragments, we generate a somewhat higher amount of short fragments than with the traditional protocol. Sequencing these very short fragments sequenced with long reads and/or paired-end sequencing increases the risk of sequencing the

adapters or generating duplicate reads, which are undesirable artifacts for ChIP-seq analysis: they can be removed but that would still mean a loss of reads and loss of information. Trimming can introduce bias in the coverage profile as well by the unbalanced removing of portions of the reads. Also, creating paired-end reads and paired-end libraries, plus the analysis of paired-end data sets, would elongate the sample-to-results time severely, including additional hands-on time and computing time, and the additional layer of complexity in the analysis is eventually pointless, as the peak caller we use (as most broad peak callers, due to the nature of the signal) cannot take advantage on paired-end information, so in the end it would not improve the peak calling. The mappability rate would also not be improved significantly, as we can already map routinely 80%–90% of the 50 bp single-end reads, and it is mainly the repeated regions where paired-end reads could really help mapping (apart from mutations), but those are rarely targeted in a ChIP-seq experiment. Another benefit of paired-end reads is the discovery of mutations, copy number variations, structural variations, but then again, these are usually not the goal of a ChIP-seq study. And there is the elevated cost as well, which we also cannot overlook. Thus, altogether we do not recommend paired-end reads because the little to no benefits are not justified by the many and significant drawbacks.

For a detailed description of the experimental setup and preparations, please see the "Materials and Methods" section of this study. Refer to Figure 1 for an overview of the reshearing method.

Data analysis pipeline and specific approach to call broad peaks. Though we were initially searching the literature for published ChIP-seq analysis pipelines and methods, eventually we developed our own analysis pipeline. In this way, we could automatically monitor the parameters that we deemed important for this iterative fragmentation study. We provide a detailed description of all the tools and operations our pipeline includes, and it is important to note that though we chose R to implement the custom parts of the analysis, it can be implemented in possibly any other platform; in fact, we would like to encourage the ChIP-seq community to come up with creative implementations, if they find our analysis pipeline useful for their projects. We chose R because of its high-level graphical capabilities and the built-in and easy-to-call statistical tests (such as correlation and covariation analyses) and because we have a vast experience with it, using it for various projects.

One of the focal points of developing a unique pipeline was the peak calling: we realized that histone marks often form broad enrichments, which have special qualities and need a special approach to call peaks properly.[22] Several ChIP-seq-related studies describe peak calling on transcription factors, which have an entirely different profile, and we also encountered publications dealing with histone marks that do not address the special needs of broad peak calling adequately (to name a few among many, Itahana et al.[34], Li and Zhou[35], and Hardy et al.[36]).

The main goal of this study was to study the effects of iterative fragmentation on the inactive histone mark H3K27me3. Inactive histone marks tend to form enrichment patterns over a broad region, and as a result, the peaks are overall lower but much wider than the ones in the enrichment profile of a transcription factor for example. Subsequently, these broad enrichments are sometimes termed islands rather than peaks, due to their size and shape.[23] These broad peaks or islands are difficult to call properly with most peak callers, because common methods such as finding peak pairs on the positive and negative strands (as used, eg, by MACS[24]) are ineffective due to the long enriched regions. Such peak callers expect well-separated pairs of peaks on the two strands, with clear summits and shoulders, where the distance between the summits of the pairs corresponds to the approximate fragment length, and the horizontal size of a peak on one strand is significantly shorter than the fragment length. Broad peaks render this method ineffective as they have no unequivocal summit, even the shoulder is difficult to identify sometimes as it often lacks a sudden drop, and their horizontal size exceeds that of the fragment, eliminating any separation between the prospective peak pairs. A few software tools such as Sicer are specifically developed to tackle this problem and call broad enrichments, disregarding the pairing of the peaks, but even they are unable to reliably identify all the enrichments, as we have observed. One of the problems is that the enrichment level is not uniform along the peak, which prompts some peak callers to call only local summits within a longer stretch of an enriched region, instead of detecting the whole enrichment as one peak. The authors of MACS demonstrate this phenomenon in a more recent publication,[37] where they show a broad enrichment of a H3K36me3 mark, which is only partially detected by MACS, broken up into countless small fragments, while other parts of the enrichment are not detected at all. (Note that since then an updated version of MACS has been released, termed MACS2, which has several options to aid proper broad peak calling; available at https://github.com/taoliu/MACS.) Another problem is the generally low enrichment, low signal-to-noise ratio, which makes defining the borders of the peaks difficult, because the enrichments often fade into the background with a smooth transition. The failure of the proper establishment of the borders also leads to the merging of some neighboring peaks: adjacent enrichments are detected as a single peak because the separating background is not recognized. Figure 2A demonstrates on an in-house generated control data set (HeLa-S3 cells, H3K27me3 histone mark) various broad peak calling issues we encountered during the test of different peak callers (Sicer, HiddenDomains, MACS, MACS2, Zinba, BroadPeak, Rseg); we also show the practical experience that led us to choose Sicer besides the fact that it often comes out in reviews[22] and benchmarks[26] as a recommended tool or a top performer when it comes to broad peak calling. Besides, this figure makes it evident that our optimized settings greatly improve peak detection.

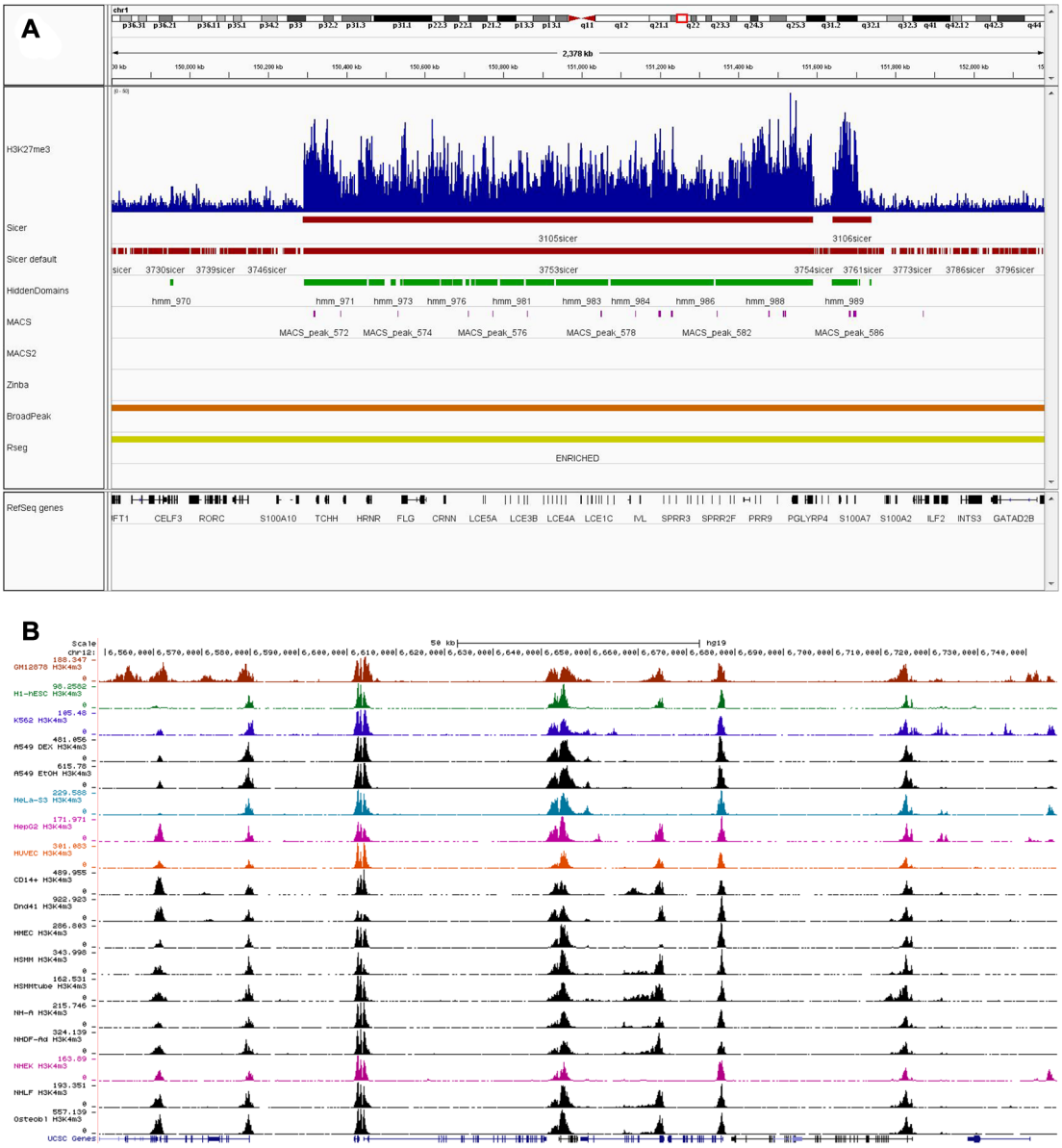

Figure 2. Comparison of the performance of various peak callers developed for broad peak detection, and the consistence of the peak type, regardless of cell types. (**A**) This image is taken from the IGV genome viewer, where various peak calling results are displayed on a HeLa-S3 in-house control data set. The upper coverage track in blue shows a longer stretch of enrichment from a ChIP-seq experiment targeting H3K27me3 histone marks. Below that the boxes show the peaks determined by selected peak callers. In a top-down order, the tracks show the peaks of the following software tools: Sicer (with the optimalized settings we established for the studied histone mark) and Sicer with default settings in red, HiddenDomains in green, MACS and MACS2 in purple, Zinba in cyan, BroadPeak in orange, and Rseg in yellow. The image represents the difficulties of detecting broad enrichments: Rseg and BroadPeak detect the whole visible region (and more) as a huge, single peak, while MACS2 and Zinba fail to recognize the enriched regions. HiddenDomains and MACS segment the enrichments into several narrow peaks. Sicer is oversensitive with the default settings, though it calls the enriched regions in the correct way, and with our optimized settings, it is able to properly differentiate between the enrichments and the background. (**B**) This image was taken from the UCSC Genome Browser, featuring 18 different cell types (or different treatments in some cases), submitted by the Broad Institute to the ENCODE project. The samples are (as they appear on the UCSC website): GM12878, H1-hESC, K562, A549 DEX, A549 EtOH, HeLa-S3, HepG2, HUVEC, CD14+, Dnd41, HMEC, HSMM, HSMMtube, NH-A, NHDF-Ad, NHEK, NHLF, and Osteobl. Apparently, the cell type does not influence what type of enrichment is generated by a certain histone mark, the enrichment types are so consistent that we can reason that the same peak caller settings should be optimal for all of them.

In this test, we compared peak callers that (with the exception of MACS, but we have put it there as it is one of the most established and widely recognized peak calling software tools) are claimed to be capable of calling broad peaks. Except for the first Sicer track, the default settings were used. If an option was available for detecting broad enrichments, or specific parameters were recommended for this type of peaks, then they were also set (eg, the option --broad for MACS2); similarly,

if there was an implemented way to improve peak detection and quality (eg, the deadzone correction for Rseg), then we used it. The results visible in the image show the erroneous detections (except for the first track) due to the inherent difficulties of detecting broad enrichments: Rseg and BroadPeak cannot determine the borders of the enrichments correctly and incorporate enriched regions and patches of background alike into a huge, single peak. MACS2 and Zinba behave in the opposite way: they see the whole region as background, failing to recognize the enriched regions; therefore, they do not call any peaks in the visible section. Hidden Domains gives one of the better results, recognizing the enriched regions, but segmenting the continuous enrichment into several narrower peaks. MACS shows an extreme case of this segmentation effect, calling only a very few of the local summits within the enrichments as narrow peaks. This figure demonstrates why we chose Sicer as our peak calling software for histone marks and also the importance of using optimized settings: Sicer was able to call the peaks relatively correctly, though by default it shows an oversensitivity, calling parts of the background as peaks too. Our optimized settings eliminated this problem, and thus Sicer managed to distinguish the enrichments from the background properly.

To tackle the problems associated with broad peak calling, we wanted to ensure that we use the appropriate software with the appropriate settings for the identification of broad peaks. As stated before, we found examples in the scientific literature where peak calling was made with uniform settings regardless of the different peak characteristics, often running the software with default settings; but our point of view is that different enrichment profiles need different peak caller settings to be detected correctly. Therefore, on our in-house validated ChIP-seq data sets of different histone marks, we tested a number of candidate peak callers and each with several different settings. Eventually, we settled on using Sicer, as we found it to be the most suitable for histone marks, and it also provides an intuitive set of parameters that provide a relatively easy way to configure the peak calling and find the optimal settings. We used the highly reliable peak set and the method we described in the "Materials and Methods" section to establish individual parameters for each histone mark. Note that we compared the peak callers in the same way (ie, monitoring the accurate detection both visually and with our selected, validated peak set). Table 1 shows the specific settings for each histone modification.

Note that we believe that the peak calling parameters should be histone mark specific (though certain histone marks yield highly similar profiles; in that case, the same settings should be suitable for those marks). There could be a significant difference among histone marks, H3K4me3 could produce enrichments along a few thousand base pairs, while H3K27me3 marks are enriched over tens or hundreds of thousands of base pairs. However, other factors, such as cell type or treatments of the cells with various chemicals, do not seem to change the enrichment type: we suggest that once we have found the optimal settings for one histone mark, the same settings can be used regardless of cell type or treatment, if the organism is the same. To support this, we used the online UCSC Genome Browser, which provides direct access to the submitted ENCODE data sets, and generated evidence showing the consistency of the enrichment type among different cell lines for the same histone mark. We selected the mark H3K4me3 and loaded all the ChIP-seq profiles that were submitted by the Broad Institute for this mark – it means 18 different data sets altogether from different cell types and different treatments. For the results, refer to Figure 2B, where it is evident that the enrichment type is not changed among the cell lines: the enrichments can become higher or lower, new peaks can appear, others can disappear, but the type of the enrichment and the general features of the peaks are always the same, and as they do not change, individual settings are not needed. The following cell lines and treatments are displayed, in default order, with their hard-coded color scheme (we provide the UCSC accession IDs in brackets): GM12878 (wgEncode EH000028), H1-hESC (wgEncodeEH000086), K562 (wgEncodeEH000048), A549 DEX (wgEncodeEH003077), A549 EtOH (wgEncodeEH003065), HeLa-S3 (wgEncode EH001017), HepG2 (wgEncodeEH000095), HUVEC (wgEncodeEH000041), CD14+ (wgEncodeEH003071), Dnd41 (wgEncodeEH002408), HMEC (wgEncodeEH 000091), HSMM (wgEncodeEH000116), HSMMtube (wgEncodeEH001007), NH-A (wgEncodeEH001032), NHDF-Ad (wgEncodeEH001053), NHEK (wgEncode EH000068), NHLF (wgEncodeEH000102), and Osteobl (wgEncodeEH003091).

Publications also support our view: those that establish different types of enrichments in ChIP-seq usually create categories based on the target protein and not on the cell line or other factors. They establish terms such as point-source peaks (typically for transcription factors), narrow peaks (for either transcription factors, in contrast with histone marks, or histone marks that generate enrichments in narrower regions, such as H3K4me3, in contrast with H3K27me3), broad enrichments (typically for histone marks), mixed signal (like what polII produces, which generates both point-source and broad peaks), enrichment islands, and diffuse or disperse marks (typically histone marks that form enrichments over tens or hundreds of thousands of base pairs, such as H3K27me3, which can be well differentiated from histone marks such as H3K4me3, which usually produce enrichments over a few thousand base pairs).[21–23,26–28]

We would also like to point out that even if the same settings can usually be applied for a histone mark across various cell types and treatment, it is always a good practice to monitor the quality of the peak calling, how faithfully can the peak caller detect the enrichments. This can be done visually, viewing the coverage graphs and the called peaks together in a genome viewer, checking if the peaks really correspond

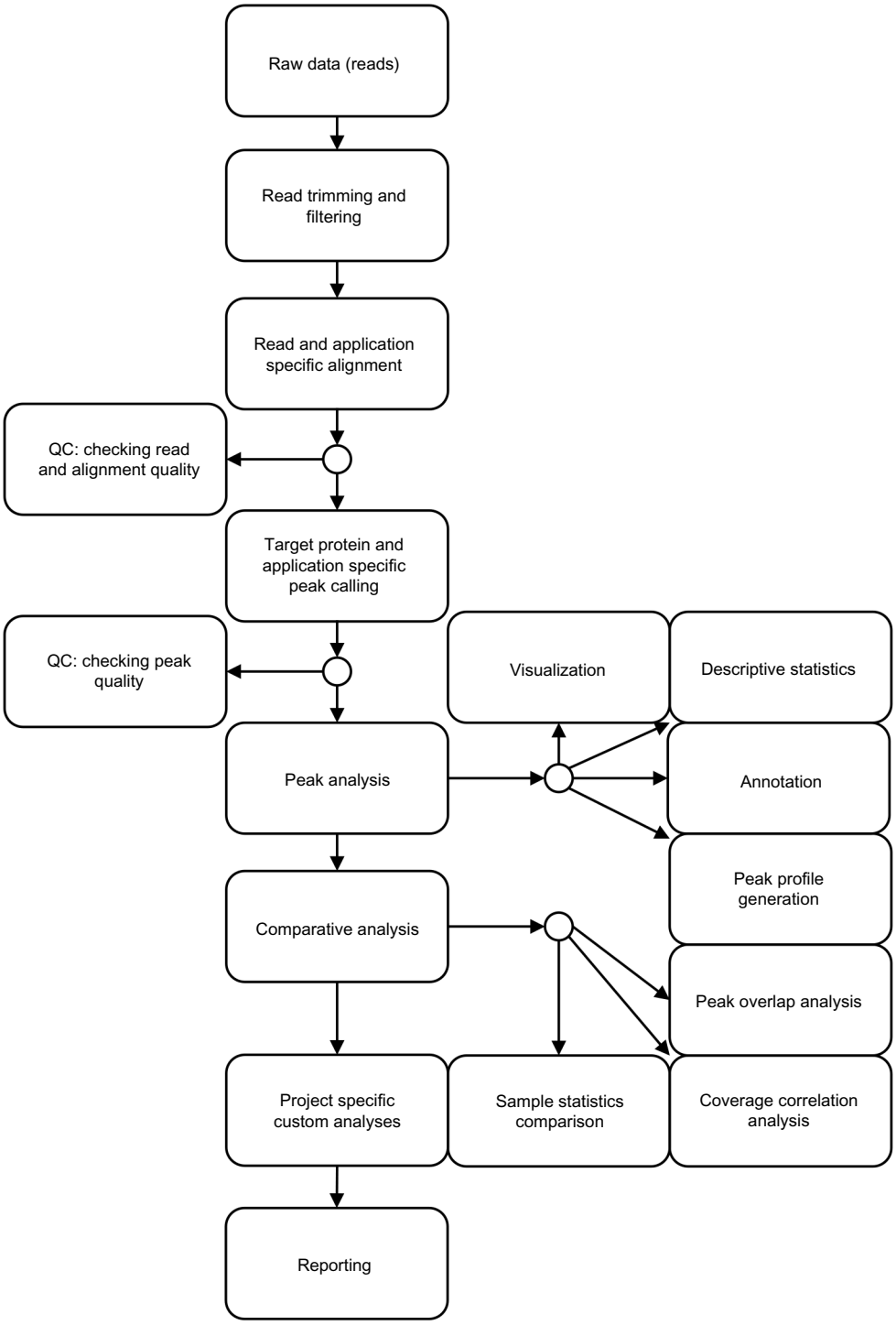

Figure 3. Flowchart of our ChIP-seq analysis pipeline. For each step, we either used a carefully selected public software or wrote our own proprietary scripts.

to enrichments, and also the peak statistics (their number, size, significance) can tell if the settings are suboptimal via, for example, a sudden and manifold drop or rise compared to previous similar samples. And it is best to reevaluate our settings when we want to apply them to another species.

Aside from the peak calling, our analysis pipeline automatically aligns the reads and performs quality control on them; then after the peak calling, it calculates a number of descriptive statistics about the peak sets to compare them, as well as annotating the peaks to see which genomic features they associate to. It also generates appropriate data sets for graphs and checks the overlaps between corresponding samples. The correlation of the data sets is examined by comparing the peak overlap fractions, as well as by analyzing the

correlation of the coverage graphs using a genome windowing method. The latter analysis provides valuable insight into correlation, covariation, and reproducibility beyond the limits of peak calling, as not every enrichment can be called as a peak and compared between samples, and when we compare the ChIP-seq results of two different methods, it is essential to also check the read accumulation and depletion in undetected regions. Therefore, our opinion is that both of these two types of analyses (comparison by peaks and without peaks) are needed for a complete evaluation. See the "Materials and Methods" section for a more detailed description. Figure 3 presents the general workflow of our data analysis pipeline.

The overlap matrices are reported in Table 2; the summary of our findings and peak statistics can be found in Table 3. The latter demonstrates the effects of resonication through various peak characteristics, such as average peak dimension or significance. Other outputs of our data analysis pipeline that demonstrates the effects of reshearing and compares them

Table 2. The overlap matrices of the three histone marks show the matching ratio of the peaks between the control and the resheared data sets, including the top 40% analysis described by the ENCODE consortium.

	CONTROL	MATCHING (%)	RESHEARED	MATCHING (%)
H3K4me1				
Total control	60635	100.00	53159	87.67
Top 40% control	24254	100.00	24233	99.91
Total resheared	47445	84.08	56426	100.00
Top 40% resheared	22454	99.49	22570	100.00
H3K4me3				
Total control	17289	100.00	17206	99.52
Top 40% control	6916	100.00	6916	100.00
Total resheared	16833	87.74	19186	100.00
Top 40% resheared	7674	100.00	7674	100.00
H3K27me3				
Total control	3044	100.00	2810	92.31
Top 40% control	1218	100.00	1216	99.84
Total resheared	2701	74.14	3643	100.00
Top 40% resheared	1431	98.22	1457	100.00

Notes: The matching of peaks is outstanding, all the top 40% ratios are way over 80%. If we compare the total peak sets, including peaks of lower significance, we still get excellent overlap ratios, although sometimes there is a larger difference between the control and resheared samples (eg, H3K27me3: 74% and 92%) – in this case, obviously one data set (the resheared in this example) has extra peaks that are not detected in the other data set, the enrichments are likely not significant enough without reshearing.

Table 3. Here the most important descriptive statistics of the peak sets are displayed for each sample.

	H3K4me1		H3K4me3		H3K27me3	
	CONTROL	RESHEARED	CONTROL	RESHEARED	CONTROL	RESHEARED
No. of peaks	60635	56426	17289	19186	3044	3643
Mean peak width	3384.76	4708.71	3353.23	4017.13	255534.48	235192.52
Mean sign. score	134.40	194.70	3061.58	3460.70	671.28	1931.41
FRiP (%)	30.69	31.81	62.07	58.63	47.79	59.95
Peaks in						
– genes (%)	16.89	20.47	54.15	50.54	62.81	62.09
– promoters (%)	10.16	10.40	63.07	56.92	53.02	53.09
– gene rich regions (%)	97.06	96.77	98.72	98.25	92.35	93.47
Pearson corr. coeff.	0.9730409		0.966679		0.9676189	
Reshearing effects	W++, M++, R+, N+		W++, M+, R+, N++		W+, S++, F++, R++, N+	

Note: Studying these statistics we can discover the relevant differences between the control and the resheared samples. From the annotation data, we highlighted three figures: how many peaks overlap with gene positions, promoter positions, and gene-rich regions. The latter is a type of quality control, as almost all enrichments are expected in the gene-rich regions. FRiP refers to the term fraction of reads in peaks from the ENCODE ChIP-seq guidelines mentioned before. The column Pearson corr. coeff. refers to the window-by-window correlation of coverages. The last column summarizes the observed effects of the reshearing on the sample.
Abbreviations: W, widening; M, merging; R, rise (in enrichment and significance); N, new peak discovery; S, separation; F, filling up (of valleys within the peak); +, observed; ++, dominant.

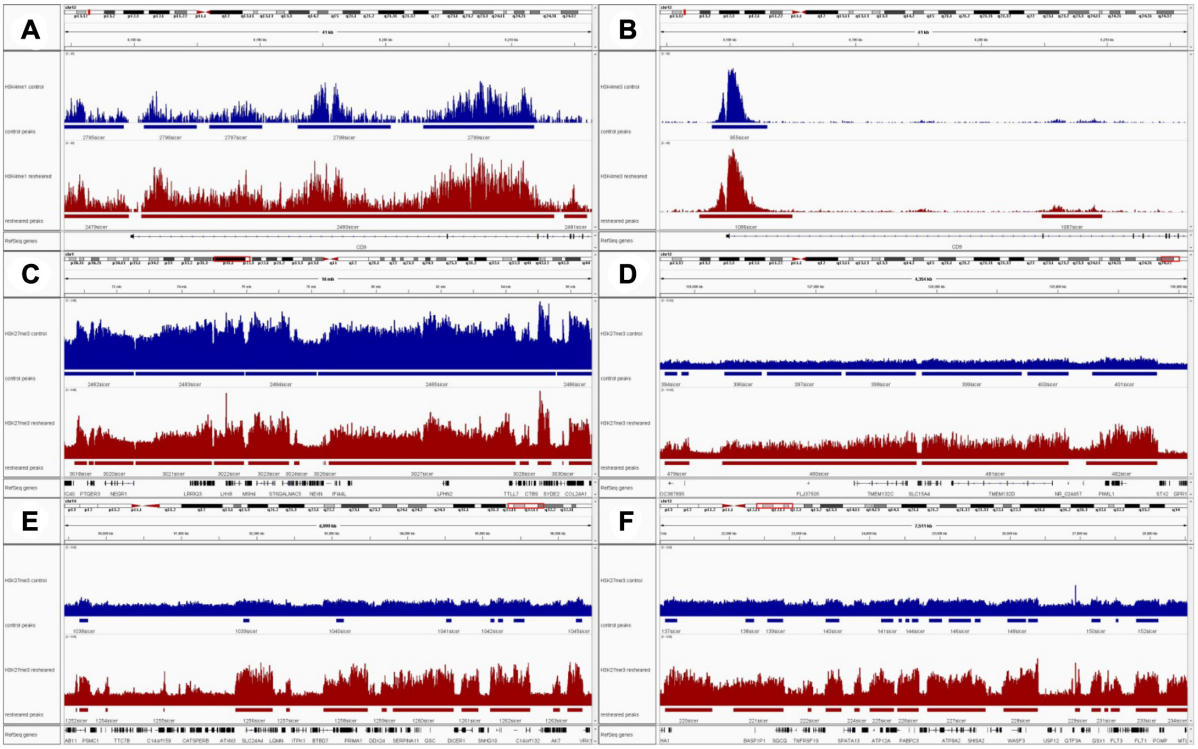

Figure 4. Various effects of the reshearing on histone marks are showcased by coverage graph and peak detection marks displayed in the IGV genome browser. The control samples are shown in blue (upper tracks), and the corresponding resheared samples are in red (lower tracks); the scales are identical for the corresponding pairs. (**A**) Reshearing effect on H3K4me1. The enrichments are visibly higher and wider in the resheared sample, but the peaks show a merging effect. (**B**) Reshearing effect on H3K4me3. Reshearing makes the peaks higher and wider, enabling the detection of smaller, otherwise insignificant peaks. (**C**–**F**) Reshearing effect on H3K27me3, which marks inactive regions and usually forms long stretches of relatively low enrichment. Note how all the pictures show a great increase in signal-to-noise ratio after reshearing. (**C**) A spectacular display of the separation effect. (**D**) The opposite of the separation effect: the peaks are dissected in the control sample but correctly recognized in the resheared sample, due to the filling-up effect. (**E**) Reshearing enables the correct peak detection by significantly increasing the signal-to-noise ratio. In the control samples, the peaks are only partially detected or not detected at all. (**F**) Several characteristic inactive mark effects are visible in this image: the control sample exhibits dissection of the peaks, partial or nondetection; reshearing eliminates these problems with the significantly better contrast to background and the filling-up effect.

across the samples are displayed in Figure 4 (A and B deal with active marks, while C–F present regions of an inactive mark) and Figure 5 (where A–C show average coverages of the control samples, D–F show average coverages of the resheared samples, and G–I show scatterplots to visualize correlations).

Inactive marks show improved sensibility and detectability. We observed that the iterative fragmentation method has a very positive effect on enrichment and peak detection for the H3K27me3 histone mark and offers a solution for the broad peak calling problems described before. Compared to the control sample, the enrichments have become substantially elevated and more significant (Table 3 and Fig. 5), and we have detected more, but narrower peaks (Table 3). The latter is caused by a separation effect: because of the significantly improved enrichments and contrast to the background, the separating background patches between adjacent peaks are better recognized, thus less merging occurs, subsequently the enrichments that are detected as merged broad peaks in the control sample often appear correctly separated in the resheared sample. In all the images in Figure 4 that deal with H3K27me3 (C–F), the greatly improved signal-to-noise ratio is apparent. In fact, reshearing has a much stronger impact on H3K27me3 than on the active marks. It appears that a significant portion (probably the majority) of the antibody-captured proteins carry long fragments that are discarded by the standard ChIP-seq method; therefore, in inactive histone mark studies, it is much more important to exploit this technique than in active mark experiments. Figure 4C showcases an example of the above-discussed separation. After reshearing, the exact borders of the peaks become recognizable for the peak caller software, while in the control sample, several enrichments are merged. Figure 4D reveals another beneficial effect: the filling up. Sometimes broad peaks contain internal valleys that cause the dissection of a single broad peak into many narrow peaks during peak detection; we can see that in the control sample, the peak borders are not recognized properly, causing the dissection of the peaks. After reshearing, we can see that in many cases, these internal valleys are filled up to a point where the broad enrichment is correctly detected as a single peak; in the displayed example, it is visible how reshearing uncovers the correct borders by filling up the valleys within the peak, resulting in the correct detection of

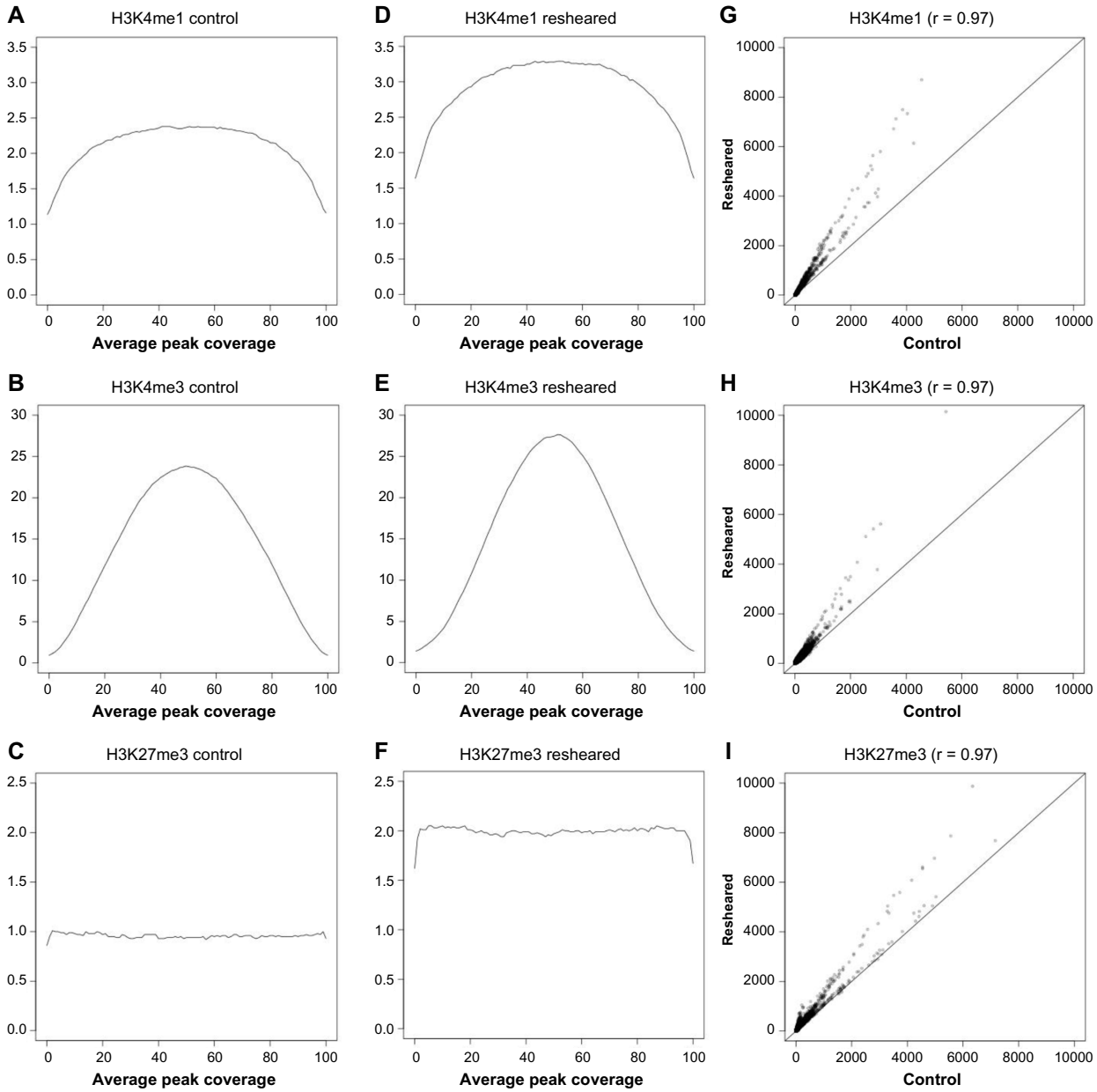

Figure 5. Average peak profiles and correlations between the resheared and control samples. The average peak coverages were calculated by binning every peak into 100 bins, then calculating the mean of coverages for each bin rank. The scatterplots show the correlation between the coverages of genomes, examined in 100 bp windows. (**A**–**C**) Average peak coverage for the control samples. The histone mark-specific differences in enrichment and characteristic peak shapes can be observed. (**D**–**F**) Average peak coverages for the resheared samples. Note that all histone marks exhibit a generally higher coverage and a more extended shoulder area. (**G**–**I**) Scatterplots show the linear correlation between the control and resheared sample coverage profiles. The distribution of markers reveals a strong linear correlation, and also some differential coverage (being preferentially higher in resheared samples) is exposed. The *r* value in brackets is the Pearson's coefficient of correlation. To improve visibility, extreme high coverage values have been removed and alpha blending was used to indicate the density of markers. This analysis provides valuable insight into correlation, covariation, and reproducibility beyond the limits of peak calling, as not every enrichment can be called as a peak, and compared between samples, and when we compare the ChIP-seq results of two different methods, it is essential to also check the read accumulation and depletion in undetected regions.

the enrichments as single continuous regions. Furthermore, due to the huge increase in the signal-to-noise ratio and the enrichment level, we were able to identify new enrichments as well in the resheared data sets: we managed to call peaks that were previously undetectable or only partially detected. Figure 4E highlights this positive impact of the increased significance of the enrichments on peak detection. Figure 4F also presents this improvement along with other positive effects that counter many typical broad peak calling problems under normal circumstances.

The immense increase in enrichments corroborate that the long fragments made accessible by iterative fragmentation are not unspecific DNA, instead they indeed carry the targeted modified histone protein H3K27me3 in this case: the

long fragments colocalize with the enrichments previously established by the traditional size selection method, instead of being distributed randomly (which would be the case if they were unspecific DNA). Evidences that the peaks and enrichment profiles of the resheared samples and the control samples are extremely closely related can be seen in Table 2, which presents the excellent overlapping ratios; Table 3, which – among others – shows a very high Pearson's coefficient of correlation close to one, indicating a high correlation of the peaks; and Figure 5, which – also among others – demonstrates the high correlation of the general enrichment profiles. If the fragments that are introduced in the analysis by the iterative resonication were unrelated to the studied histone marks, they would either form new peaks, decreasing the overlap ratios significantly, or distribute randomly, raising the level of noise, reducing the significance scores of the peak. Instead, we observed very consistent peak sets and coverage profiles with high overlap ratios and strong linear correlations, and also the significance of the peaks was improved, and the enrichments became higher compared to the noise; that is how we can conclude that the longer fragments introduced by the refragmentation are indeed belong to the studied histone mark, and they carried the targeted modified histones. In fact, the rise in significance is so high that we arrived at the conclusion that in case of such inactive marks, the majority of the modified histones could be found on longer DNA fragments. The improvement of the signal-to-noise ratio and the peak detection is significantly greater than in the case of active marks (see below, and also in Table 3); therefore, it is essential for inactive marks to utilize reshearing to enable proper analysis and to prevent losing valuable information.

Active marks exhibit higher enrichment, higher background. Reshearing clearly affects active histone marks as well: even though the increase of enrichments is less, similarly to inactive histone marks, the resonicated longer fragments can enhance peak detectability and signal-to-noise ratio. This is well represented by the H3K4me3 data set, where we detect more peaks compared to the control. These peaks are higher, wider, and have a larger significance score in general (Table 3 and Fig. 5). We found that refragmentation undoubtedly increases sensitivity, as some smaller peaks that were unidentifiable for the peak caller in the control data set become detectable with reshearing. These smaller peaks, however, usually appear out of gene and promoter regions; therefore, we conclude that they have a higher chance of being false positives, knowing that the H3K4me3 histone modification is strongly associated with active genes.[38] Another evidence that makes it certain that not all the extra fragments are valuable is the fact that the ratio of reads in peaks is lower for the resheared H3K4me3 sample, showing that the noise level has become slightly higher. Nonetheless, this is compensated by the even higher enrichments, leading to the overall better significance scores of the peaks despite the elevated background.

We also observed that the peaks in the refragmented sample have an extended shoulder area (that is why the peaks have become wider), which is again explicable by the fact that iterative sonication introduces the longer fragments into the analysis, which would have been discarded by the conventional ChIP-seq method, which does not involve the long fragments in the sequencing and subsequently the analysis. The detected enrichments extend sideways, which has a detrimental effect: sometimes it causes nearby separate peaks to be detected as a single peak. This is the opposite of the separation effect that we observed with broad inactive marks, where reshearing helped the separation of peaks in certain cases.

The H3K4me1 mark tends to produce significantly more and smaller enrichments than H3K4me3, and many of them are situated close to each other. Therefore – while the aforementioned effects are also present, such as the increased size and significance of the peaks – this data set showcases the merging effect extensively: nearby peaks are detected as one, because the extended shoulders fill up the separating gaps. H3K4me3 peaks are higher, more discernible from the background and from each other, so the individual enrichments usually remain well detectable even with the reshearing method, the merging of peaks is less frequent. With the more numerous, quite smaller peaks of H3K4me1 however the merging effect is so prevalent that the resheared sample has less detected peaks than the control sample. As a consequence after refragmenting the H3K4me1 fragments, the average peak width broadened significantly more than in the case of H3K4me3, and the ratio of reads in peaks also increased instead of decreasing. This is because the regions between neighboring peaks have become integrated into the extended, merged peak region. Table 3 describes the general peak characteristics and their changes mentioned above.

Figure 4A and B highlights the effects we observed on active marks, such as the generally higher enrichments, as well as the extension of the peak shoulders and subsequent merging of the peaks if they are close to each other. Figure 4A shows the reshearing effect on H3K4me1. The enrichments are visibly higher and wider in the resheared sample, their increased size means better detectability, but as H3K4me1 peaks often occur close to each other, the widened peaks connect and they are detected as a single joint peak. Figure 4B presents the reshearing effect on H3K4me3. This well-studied mark usually indicating active gene transcription forms already significant enrichments (typically higher than H3K4me1), but reshearing makes the peaks even higher and wider. This has a positive effect on small peaks: these mark rare histone modification profiles, which only occur in the minority of the studied cells, but with the increased sensitivity of reshearing these "hidden" peaks become detectable by accumulating a larger mass of reads.

Discussion

In this study, we demonstrated the effects of iterative fragmentation, a method that involves the resonication of DNA fragments after ChIP. Additional rounds of shearing without size selection allow longer fragments to be included

in the analysis, which are usually discarded before sequencing with the traditional size selection method. In the course of this study, we examined histone marks that produce wide enrichment islands (H3K27me3), as well as ones that generate narrow, point-source enrichments (H3K4me1 and H3K4me3). We have also developed a bioinformatics analysis pipeline to characterize ChIP-seq data sets prepared with this novel method and suggested and described the use of a histone mark-specific peak calling procedure.

Among the histone marks we studied, H3K27me3 is of particular interest as it indicates inactive genomic regions, where genes are not transcribed, and therefore, they are made inaccessible with a tightly packed chromatin structure, which in turn is more resistant to physical breaking forces, like the shearing effect of ultrasonication. Thus, such regions are much more likely to produce longer fragments when sonicated, for example, in a ChIP-seq protocol; therefore, it is essential to involve these fragments in the analysis when these inactive marks are studied.

The iterative sonication method increases the number of captured fragments available for sequencing: as we have observed in our ChIP-seq experiments, this is universally true for both inactive and active histone marks; the enrichments become larger and more distinguishable from the background. The fact that these longer extra fragments, which would be discarded with the conventional method (single shearing followed by size selection), are detected in previously confirmed enrichment sites proves that they indeed belong to the target protein, they are not unspecific artifacts, a significant population of them contains valuable information. This is particularly true for the long enrichment forming inactive marks such as H3K27me3, where a great portion of the target histone modification can be found on these large fragments.

An unequivocal effect of the iterative fragmentation is the increased sensitivity: peaks become higher, more significant, previously undetectable ones become detectable. However, as it is often the case, there is a trade-off between sensitivity and specificity: with iterative refragmentation, some of the newly emerging peaks are quite possibly false positives, because we observed that their contrast with the usually higher noise level is often low, subsequently they are predominantly accompanied by a low significance score, and several of them are not confirmed by the annotation. Besides the raised sensitivity, there are other salient effects: peaks can become wider as the shoulder region becomes more emphasized, and smaller gaps and valleys can be filled up, either between peaks or within a peak. The effect is largely dependent on the characteristic enrichment profile of the histone mark. The former effect (filling up of inter-peak gaps) is frequently occurring in samples where many smaller (both in width and height) peaks are in close vicinity of each other, such as in the H3K4me1 data set. With such a peak profile the extended and subsequently overlapping shoulder regions can hamper proper peak detection, causing the perceived merging of peaks that should be separate. Narrow peaks that are already very significant and isolated (eg, H3K4me3) are less affected. The other type of filling up, occurring in the valleys within a peak, has a considerable effect on marks that produce very broad, but generally low and variable enrichment islands (eg, H3K27me3). This phenomenon can be very positive, because while the gaps between the peaks become more recognizable, the widening effect has much less impact, given that the enrichments are already very wide; hence, the gain in the shoulder area is insignificant compared to the total width. In this way, the enriched regions can become more significant and more distinguishable from the noise and from one another.

Literature search revealed another noteworthy ChIP-seq protocol that affects fragment length and thus peak characteristics and detectability: ChIP-exo.[39] This protocol employs a lambda exonuclease enzyme to degrade the double-stranded DNA unbound by proteins. We tested ChIP-exo in a separate scientific project to see how it affects sensitivity and specificity, and the comparison came naturally with the iterative fragmentation method. The effects of the two methods are shown in Figure 6 comparatively, both on point-source peaks and on broad enrichment islands. According to our experience ChIP-exo is almost the exact opposite of iterative fragmentation, regarding effects on enrichments and peak detection. As written in the publication of the ChIP-exo method, the specificity is enhanced, false peaks are eliminated, but some real peaks also disappear, probably due to the exonuclease enzyme failing to properly stop digesting the DNA in certain cases. Therefore, the sensitivity is generally decreased. On the other hand, the peaks in the ChIP-exo data set have universally become shorter and narrower, and an improved separation is attained for marks where the peaks occur close to each other. These effects are prominent when the studied protein generates narrow peaks, such as transcription factors, and certain histone marks, for example, H3K4me3. However, if we apply the techniques to experiments where broad enrichments are generated, which is characteristic of certain inactive histone marks, such as H3K27me3, then we can observe that broad peaks are less affected, and rather affected negatively, as the enrichments become less significant; also the local valleys and summits within an enrichment island are emphasized, promoting a segmentation effect during peak detection, that is, detecting the single enrichment as several narrow peaks.

As a resource to the scientific community, we summarized the effects for each histone mark we tested in the last row of Table 3. The meaning of the symbols in the table: W = widening, M = merging, R = rise (in enrichment and significance), N = new peak discovery, S = separation, F = filling up (of valleys within the peak); + = observed, and ++ = dominant. Effects with one + are usually suppressed by the ++ effects, for example, H3K27me3 marks also become wider (W+), but the separation effect is so prevalent (S++) that the average peak width eventually becomes shorter, as large peaks are being split. Similarly, merging H3K4me3 peaks are present (M+), but new peaks emerge in great numbers (N++) with the rise

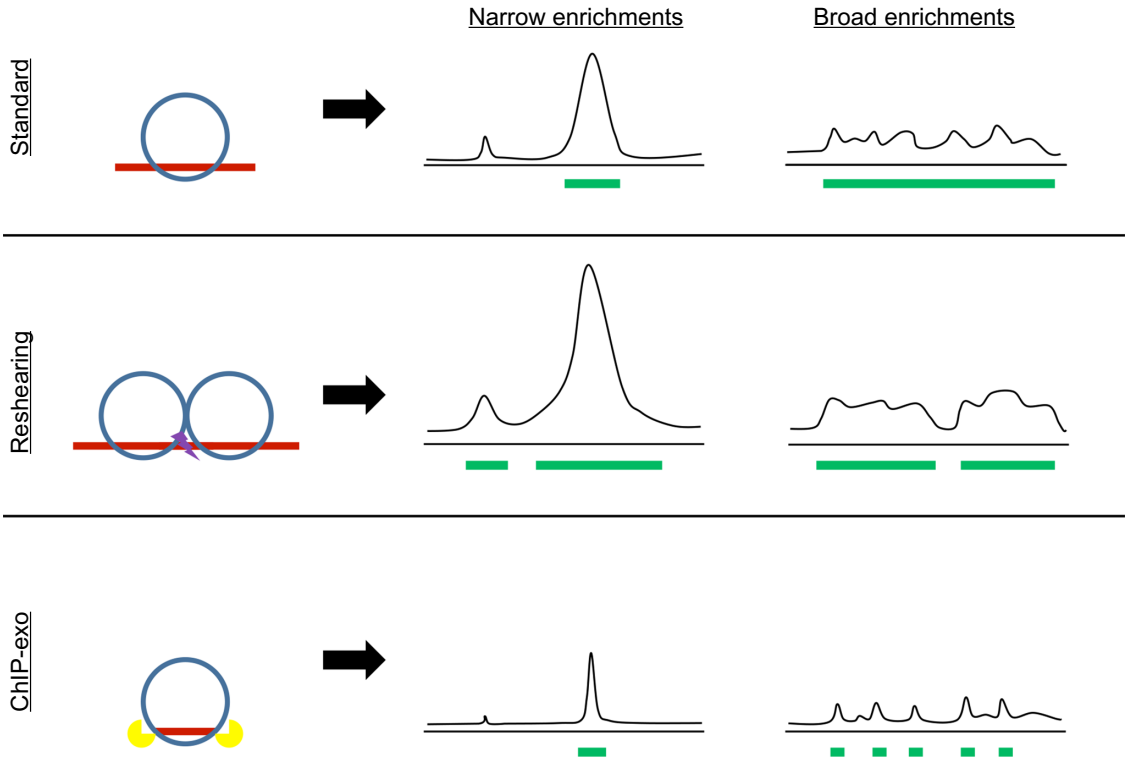

Figure 6. Schematic summarization of the effects of ChIP-seq enhancement techniques. We compared the reshearing technique that we use to the ChIP-exo technique. The blue circle represents the protein, the red line represents the DNA fragment, the purple lightning refers to sonication, and the yellow symbol is the exonuclease. On the right example, coverage graphs are displayed, with a likely peak detection pattern (detected peaks are shown as green boxes below the coverage graphs). In contrast with the standard protocol, the reshearing technique incorporates longer fragments in the analysis through additional rounds of sonication, which would otherwise be discarded, while ChIP-exo decreases the size of the fragments by digesting the parts of the DNA not bound to a protein with lambda exonuclease. For profiles consisting of narrow peaks, the reshearing technique increases sensitivity with the more fragments involved; thus, even smaller enrichments become detectable, but the peaks also become wider, to the point of being merged. ChIP-exo, on the other hand, decreases the enrichments, some smaller peaks can disappear altogether, but it increases specificity and enables the accurate detection of binding sites. With broad peak profiles, however, we can observe that the standard technique often hampers proper peak detection, as the enrichments are only partial and difficult to distinguish from the background, due to the sample loss. Therefore, broad enrichments, with their typical variable height is often detected only partially, dissecting the enrichment into several smaller parts that reflect local higher coverage within the enrichment or the peak caller is unable to differentiate the enrichment from the background properly, and consequently, either several enrichments are detected as one, or the enrichment is not detected at all. Reshearing improves peak calling by filling up the valleys within an enrichment and causing better peak separation. ChIP-exo, however, promotes the partial, dissecting peak detection by deepening the valleys within an enrichment. In turn, it can be utilized to determine the locations of nucleosomes with precision.

of significance; thus, eventually the total peak number will be increased, instead of decreased (as for H3K4me1).

The following recommendations are only general ones, specific applications might demand a different approach, but we believe that the iterative fragmentation effect is dependent on two factors: the chromatin structure and the enrichment type, that is, whether the studied histone mark is found in euchromatin or heterochromatin and whether the enrichments form point-source peaks or broad islands. Therefore, we expect that inactive marks that produce broad enrichments such as H4K20me3 should be similarly affected as H3K27me3 fragments, while active marks that generate point-source peaks such as H3K27ac or H3K9ac should give results similar to H3K4me1 and H3K4me3. In the future, we plan to extend our iterative fragmentation tests to encompass more histone marks, including the active mark H3K36me3, which tends to generate broad enrichments and evaluate the effects.

Implementation of the iterative fragmentation technique would be beneficial in scenarios where increased sensitivity is required, more specifically, where sensitivity is favored at the cost of reduced specificity. Such applications include ChIP-seq from limited biological material (eg, forensic, ancient, or biopsy samples) or where the study is limited to known enrichment sites, therefore the presence of false peaks is indifferent (eg, comparing the enrichment levels quantitatively in samples of cancer patients, using only selected, verified enrichment sites over oncogenic regions).

On the other hand, we would caution against using iterative fragmentation in studies for which specificity is more important than sensitivity, for example, de novo peak discovery, identification of the exact location of binding sites, or biomarker research. For such applications, other methods such as the aforementioned ChIP-exo are more appropriate.

The advantage of the iterative refragmentation method is also indisputable in cases where longer fragments tend to carry the regions of interest, for example, in studies of heterochromatin or genomes with extremely high GC content, which are more resistant to physical fracturing.

Conclusion

The effects of iterative fragmentation are not universal; they are largely application dependent: whether it is beneficial or detrimental (or possibly neutral) is determined by the histone mark in question and the objectives of the study. In this study, we have described its effects on multiple histone marks with the intention of offering guidance to the scientific community, shedding light on the effects of reshearing and their connection to different histone marks, facilitating informed decision making regarding the application of iterative fragmentation in different research scenarios.

Acknowledgment

The authors would like to extend their gratitude to Vincent Botta for his expert advices and his help with image manipulation.

Author Contributions

All the authors contributed substantially to this work. ML wrote the manuscript, designed the analysis pipeline, performed the analyses, interpreted the results, and provided technical assistance to the ChIP-seq sample preparations. JH designed the refragmentation method and performed the ChIPs and the library preparations. A-CV performed the shearing, including the refragmentations, and she took part in the library preparations. MT maintained and provided the cell cultures and prepared the samples for ChIP. SM wrote the manuscript, implemented and tested the analysis pipeline, and performed the analyses. DP coordinated the project and assured technical assistance. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Marino-Ramirez L, Kann MG, Shoemaker BA, Landsman D. Histone structure and nucleosome stability. *Expert Rev Proteomics*. 2005;2(5):719–29.
2. Oike T, Ogiwara H, Amornwichet N, Nakano T, Kohno T. Chromatin-regulating proteins as targets for cancer therapy. *J Radiat Res*. 2014;55(4):613–28.
3. Baylin SB, Jones PA. A decade of exploring the cancer epigenome – biological and translational implications. *Nat Rev Cancer*. 2011;11(10):726–34.
4. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell*. 2012;150(1):12–27.
5. Hendrich B, Bickmore W. Human diseases with underlying defects in chromatin structure and modification. *Hum Mol Genet*. 2001;10(20):2233–42.
6. Ho JW, Bishop E, Karchenko PV, Negre N, White KP, Park PJ. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*. 2011;12:134.
7. Fanelli M, Amatori S, Barozzi I, Minucci S. Chromatin immunoprecipitation and high-throughput sequencing from paraffin-embedded pathology tissue. *Nat Protoc*. 2011;6(12):1905–19.
8. Gilfillan GD, Hughes T, Sheng Y, et al. Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*. 2012;13:645.
9. Shankaranarayanan P, Mendoza-Parra MA, Walia M, et al. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat Methods*. 2011;8(7):565–7.
10. Kasoji SK, Pattenden SG, Malc EP, et al. Cavitation enhancing nanodroplets mediate efficient DNA fragmentation in a bench top ultrasonic water bath. *PLoS One*. 2015;10(7):e0133014.
11. Mokry M, Hatzis P, de Bruijn E, et al. Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PLoS One*. 2010;5(11):e15092.
12. Schwartz YB, Pirrotta V. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet*. 2007;8(1):9–22.
13. Wang J, Lawry ST, Cohen AL, Jia S. Chromosome boundary elements and regulation of heterochromatin spreading. *Cell Mol Life Sci*. 2014;71(24):4841–52.
14. Teytelman L, Ozaydin B, Zill O, et al. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One*. 2009;4(8):e6700.
15. Auerbach RK, Euskirchen G, Rozowsky J, et al. Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A*. 2009;106(35):14926–31.
16. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet*. 2014;15(11):709–21.
17. Gentsch GE, Smith JC. Investigating physical chromatin associations across the Xenopus genome by chromatin immunoprecipitation. *Cold Spring Harb Protoc*. 2014;2014(5). doi: 10.1101/pdb.prot080614.
18. Barrilleaux BL, Cotterman R, Knoepfler PS. Chromatin immunoprecipitation assays for Myc and N-Myc. *Methods Mol Biol*. 2013;1012:117–33.
19. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17(1):10–12.
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
21. Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 2012;22(9):1813–31.
22. Bailey T, Krajewski P, Ladunga I, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol*. 2013;9(11):e1003326.
23. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. 2009;25(15):1952–8.
24. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
25. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol*. 2011;12(7):R67.
26. Starmer J, Magnuson T. Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. *BMC Bioinformatics*. 2016;17:144.
27. Wang J, Lunyak VV, Jordan IK. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics*. 2013;29(4):492–3.
28. Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*. 2011;27(6):870–1.
29. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178–92.
30. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime CIS-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576–89.
31. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: 2013.
32. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
33. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
34. Itahana Y, Zhang J, Goke J, et al. Histone modifications and p53 binding poise the p21 promoter for activation in human embryonic stem cells. *Sci Rep*. 2016;6:28112.
35. Li G, Zhou L. Genome-wide identification of chromatin transitional regions reveals diverse mechanisms defining the boundary of facultative heterochromatin. *PLoS One*. 2013;8(6):e67156.
36. Hardy K, Wu F, Tu W, et al. Identification of chromatin accessibility domains in human breast cancer stem cells. *Nucleus*. 2016;7(1):50–67.
37. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*. 2012;7(9):1728–40.
38. Martin C, Zhang Y. The diverse functions of histone lysine methylation. *Nat Rev Mol Cell Biol*. 2005;6(11):838–49.
39. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*. 2011;147(6):1408–19.